Open Science at LHCb

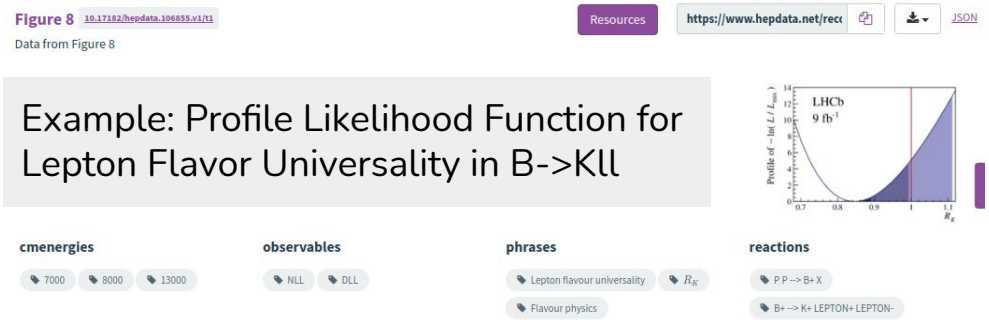
Sebastian Neubert[1]

on behalf of the LHCb Collaboration

[1]**HISKP Bonn**

CERN Open Science Strategy Working Group,
20.09.2021

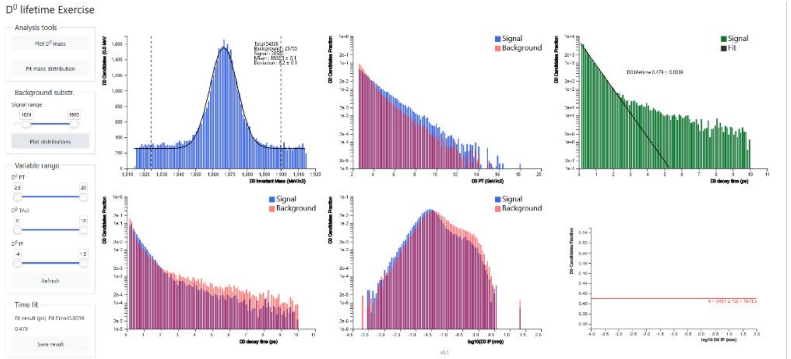
Open Science Activities



Open Data

- Level 1: 75 HEPData Entries
- Level 2: Jet samples for data science (MC)
- Level 3: Run 1 open data release planned for end 2021

Outreach and Education



- LHCb Masterclass
  - measuring the lifetime of the D0 meson
  - new Masterclass on spectroscopy in preparation
- Matter Antimatter Differences - University level course

Open data policy and mission

Important quotes from the LHC Open Data Policy

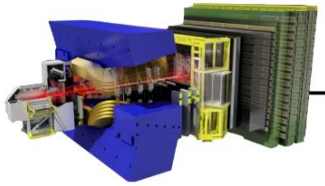
"*Making data available responsibly (applying FAIR standards), at different levels of abstraction and at different points in time, allows* ***the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community***"

L1: "*To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,4 with selection routines stored in specialised tools.*"

L2: "*The information provided will be sufficient to allow* ***high-quality analysis of the data*** *including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A* ***limited level of support*** *for users of the Level 3 Open Data will be provided on a* ***best-effort basis by the collaborations.***"

LHCb Level 3 Data

Release policy: 50% @ 5yrs, 100% @ 10yrs
after end of running period



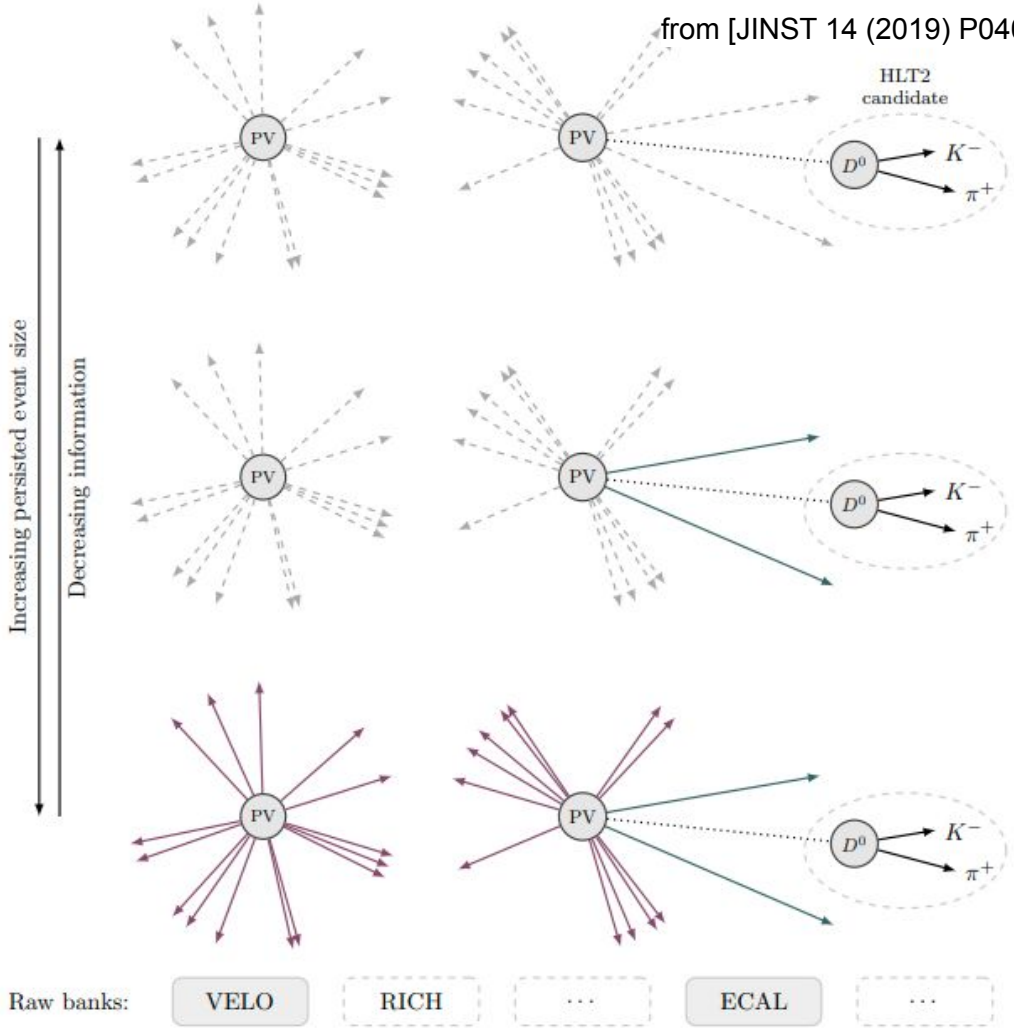
- Level 3 data in LHCb **defined as the output of the stripping**
- Same level of abstraction accessed by LHCb members
- Contains **comprehensive set of selections (1620 selections in v21)**
- Organized in ~10 streams, according to physics signature
- Storage: **~0.9 PB for Run I** (2011/12 run period) will rise to ~10PB with Run 2
- Software needed to access data (DaVinci) is open source, available via CVMFS (or container)
- Documentation on open data portal needs updating. LHCb Starterkit openly available

Current and future developments

- **Fine grained control over which event details** are written to storage, depending on physics
- **Since Run 2: Turbo stream** (using online reconstruction)

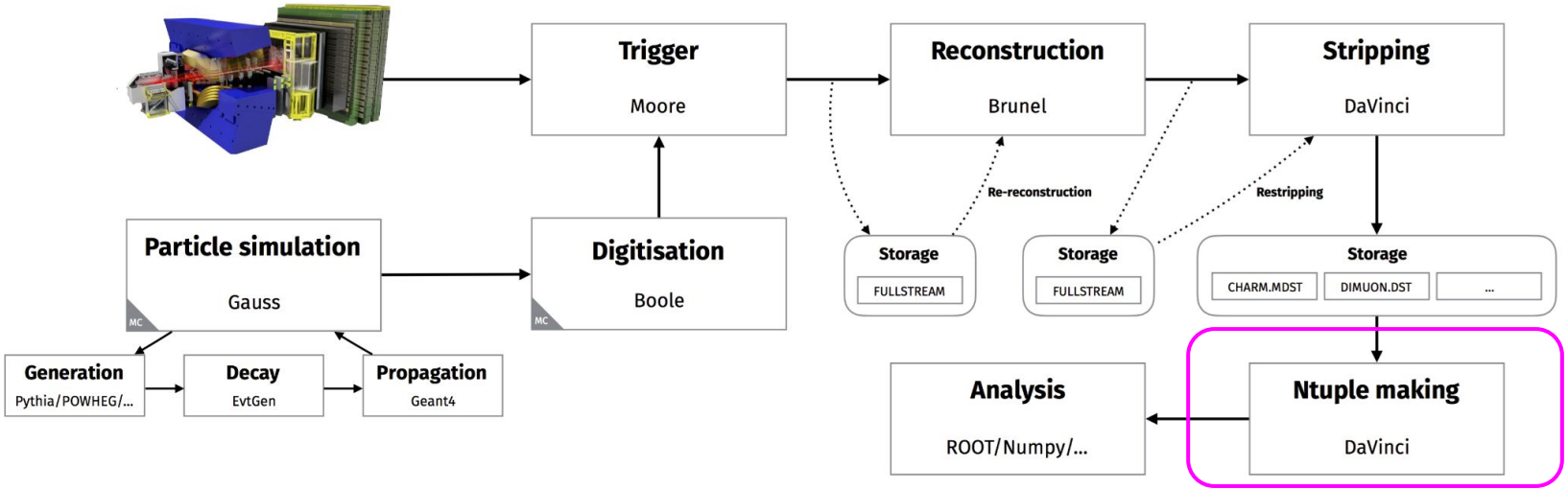
- **Run 3: Real Time Analysis** Full reconstruction online
- **Attention will shift towards provenance tracking**

from [JINST 14 (2019) P04006]



LHCb Open Data beyond Run 2

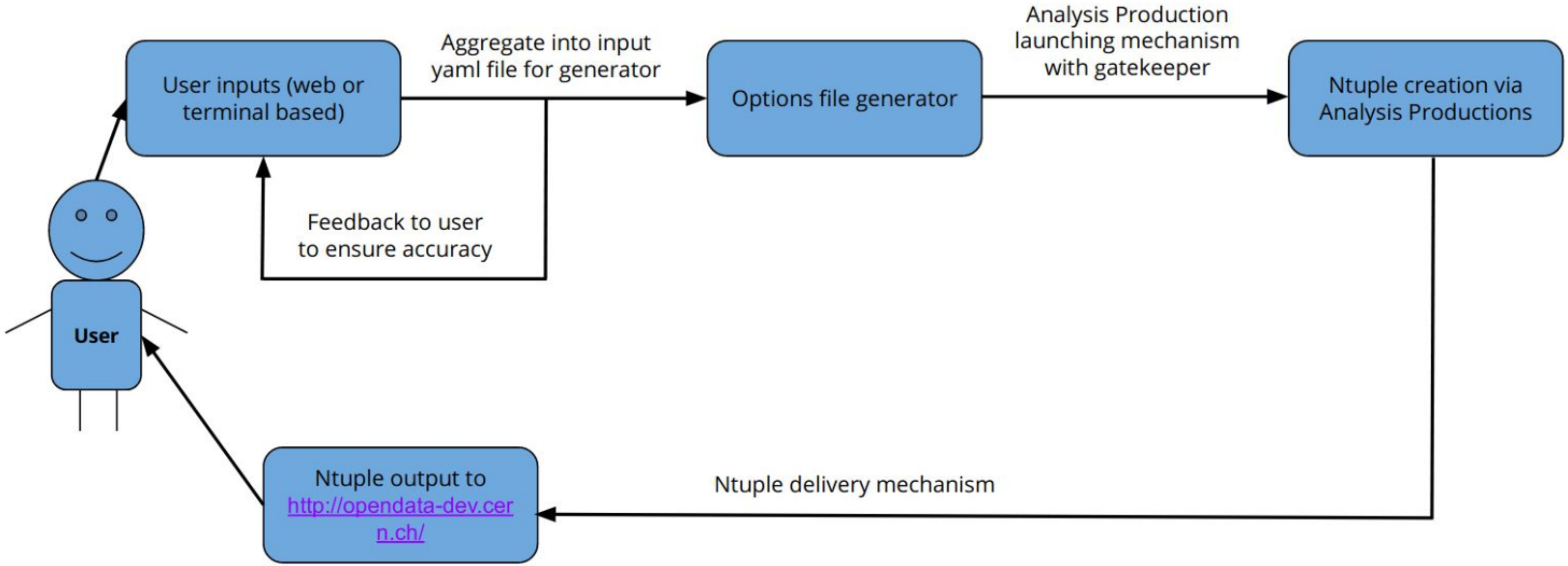
- LHCb Level 3 data **one order of magnitude more volume than other LHC experiments**, due to “signal-rich” physics (needed for precision measurements)
- Creating replicas just for the open data portal seems not efficient



Instead: Provide direct access to data on the grid

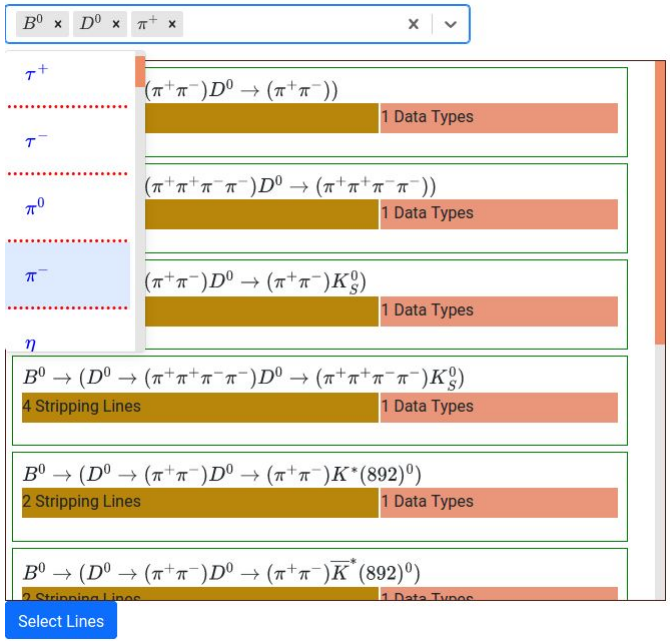
The NTuple Wizard

- NTuple Wizard: provide **secure interface** to access our replicas
- Further advantages: fine grained access control, output files easier to handle (no LHCb software needed)

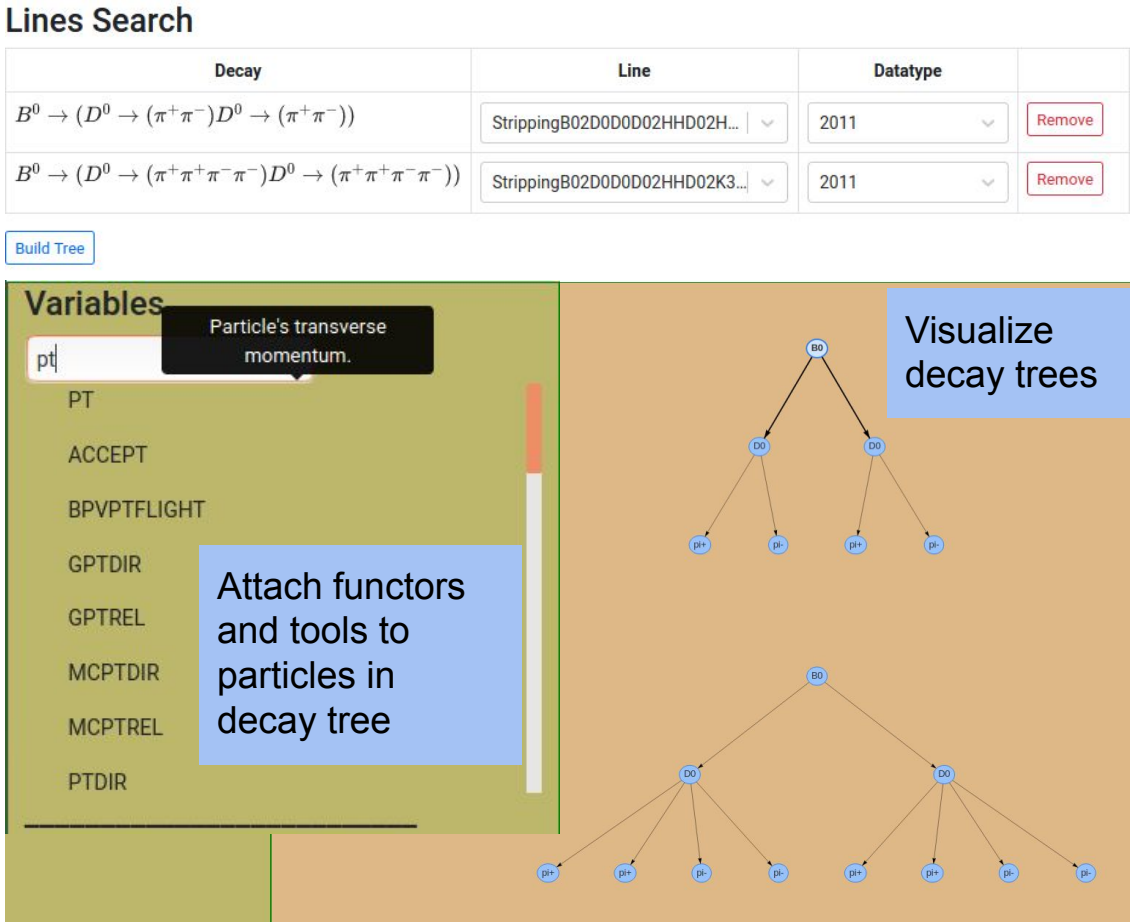


Sneak Peak: NTuple Wizard (work in progress)

Webportal offers dedicated search for processes selected in LHCb data:



Configured job launched into LHCb production system. Results will be delivered to open data portal

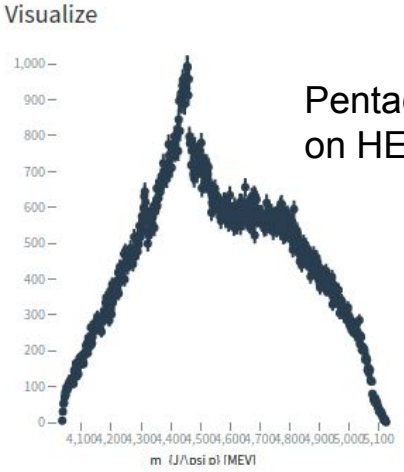


Analysis Preservation

https://doi.org/10.17182/hepdata.89271

As part of the publication procedure LHCb analysis teams are required to:

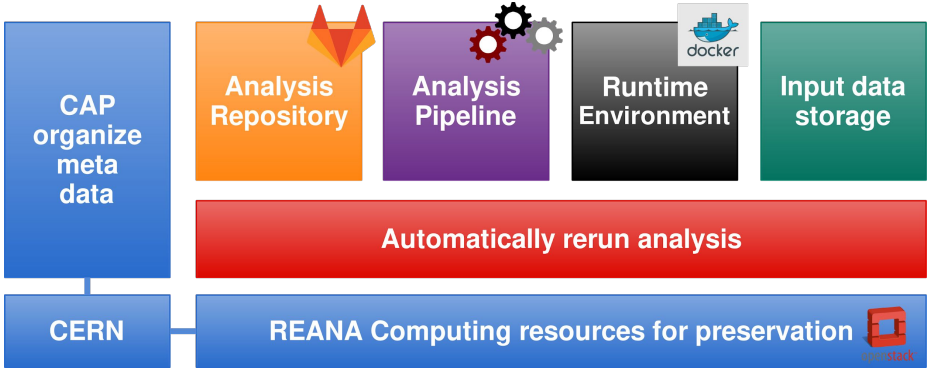
- archive ntuples on EOS or grid storage
- Analysis code on gitlab.cern.ch handed over to physics working group



Pentaquark on HEPData

Where useful,
processed analysis products are made available through **HEPdata** .

Possible future developments:
Analysis software for publication requires high level of automation



Summary

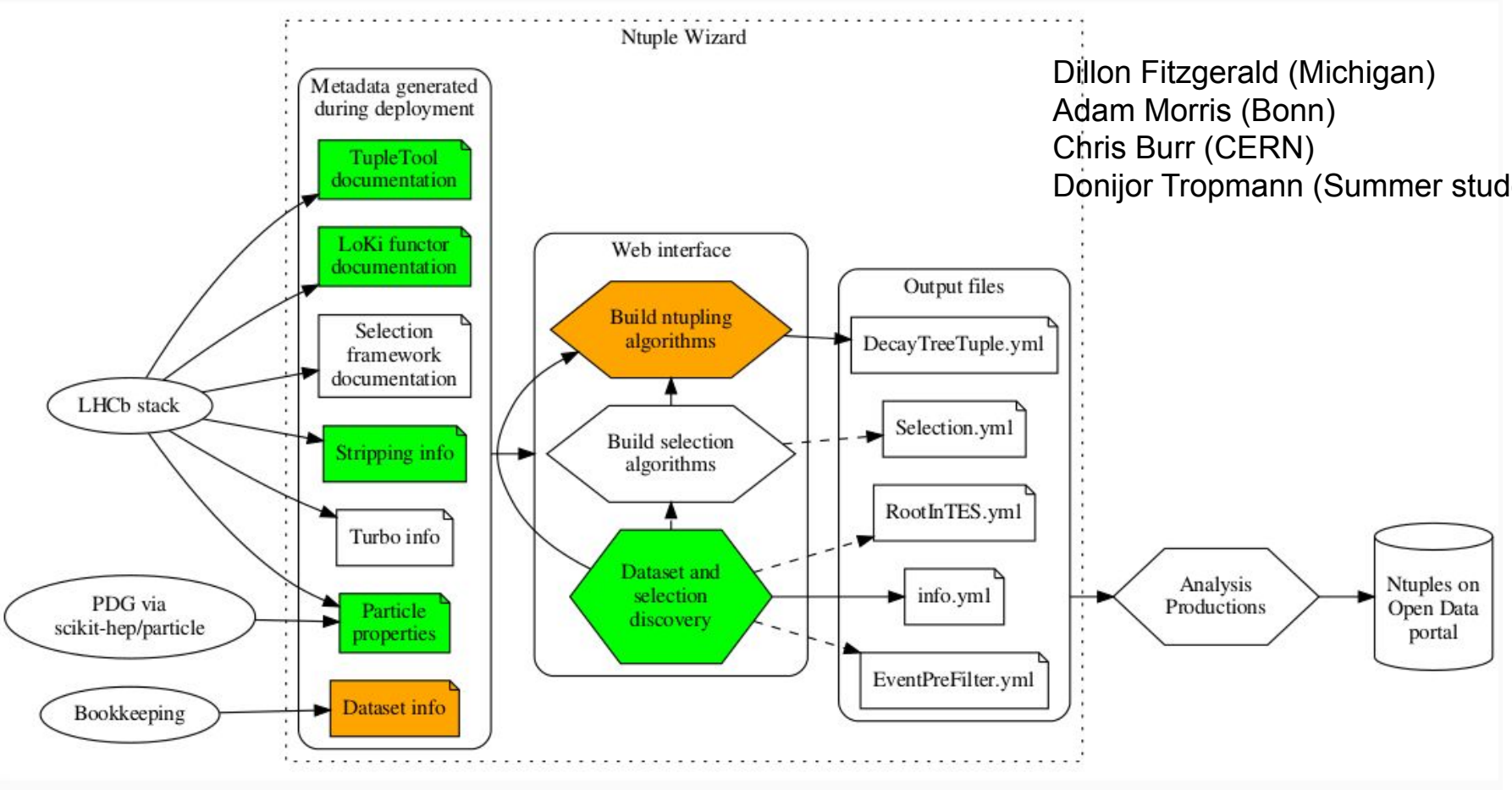
- **Open Science in LHCb focused on Open Data and Outreach/Education**

- Open data: **preparing release of Run I** stripping output this Year

- Making large datasets accessible (not just available):
  NTuple Wizard will enable barrier-free access by 3rd parties
  - Word of caution: We strongly believe in the willingness of the public to handle the data responsibly.

- Analysis preservation:
  - All analysis code and intermediate results archived internally
  - Publishing code only makes sense if it runs and can be reused

NTuple Wizard - details



Dillon Fitzgerald (Michigan)
Adam Morris (Bonn)
Chris Burr (CERN)
Donijor Tropmann (Summer student)